Frutos-López, M., Medina-Chanca, H., Sanz-Rodríguez, S., Peláez-Moreno, C. & Díaz-de-María, F. (2012). Perceptually-aware bilateral filtering for quality improvement in low bit rate video coding. In Domanski, M., et al. (eds.). *2012 Picture Coding Symposium PCS 2012: Proceedings*. (pp. 477-480). IEEE.
DOI: /http://dx.doi.org/10.1109/PCS.2012.6213258

Perceptually-Aware Bilateral Filtering for Quality Improvement in Low Bit Rate Video Coding

Manuel de-Frutos-López, Helen Medina-Chanca, Sergio Sanz-Rodríguez, Carmen Peláez-Moreno and Fernando Díaz-de-María
Department of Signal Theory and Communications. University Carlos III Madrid, Spain
e-mail: (see http://www.gpm.tsc.uc3m.es/)

***Abstract*—Perceptual coding has become of great interest in modern video coding due to the need for higher compression rates. Many previous works have been carried out to incorporate perceptual information to hybrid video encoders, either modifying the quantization parameter according to a certain perceptual resource allocation map or preprocessing video sequences for removing information that is not perceptually relevant. The first strategy is limited by the presence of blocking artifacts and the second one lacks of adaptation to video content. In this paper, a novel and simple approach is proposed, which performs a smart filtering prior to the encoding process preserving both the structural and motion information. The experiments prove that the use of proposed method implemented on an H.264 encoder significantly improves its perceptual quality for low bit rates.**

I. Introduction

Recently, perceptual video coding has become an important research area due to the constant need for higher compression rates. Although the bandwidth availability seems to be continuously increasing, new application scenarios keep coming up and testing the limits of current video coding standards. Besides High Definition content, the challenge of ubiquitous access to video content from portable and handheld devices makes high compression rate video coding systems of paramount importance.

Apart from the adoption of non-uniform quantization matrices in block-based discrete cosine transform codecs, perceptual techniques have not found their place yet. These techniques aim to allocate fewer resources (bits) to those areas in which the distortion would be less noticeable, saving bits for other regions more sensitive to distortion. The techniques for estimating the distortion sensitivity typically focus on two features of the Human Visual System (HVS): perceptual masking and visual attention.

Perceptual masking relates to those limitations in HVS that prevent from perceiving certain details in the presence of more intense stimuli. Although the related masking phenomena are not completely understood, some authors classify each frame region according to its texture and brightness in order to determine how these features affect perception. In [1], for example, the blocks are classified into dark, bright, smooth, chaotic, and detailed in order to find an optimum resource allocation, accepting that the HVS is less sensitive to distortion in those areas with very high or very low luminance levels or complex textures, while it proves more sensitive in presence of large structural details, such as sharp edges. A similar approach is adopted in [2] to distinguish between random and structured textures.

Visual attention is also a key factor in visual sensitivity due to the so-called foveation phenomenon, which relays on the fact that visual receptors are not equally distributed on the retina and a small area around the gaze point (fovea) concentrates the highest receptors density. According to this, visual acuity is maximal near fixation point and it progressively decays with eccentricity. Some approaches, such as that in [3], distinguish low-priority regions simulating foveation behavior in order to encode them more coarsely, saving bits for high-priority regions.

Besides the specific applications in which eye tracking is enabled, fixation point estimation in general-purpose non-interactive video is still an open research problem. The variety of proposed solutions can be grouped into two main work lines: top-down and bottom-up strategies. On the one hand, it is not possible to apply top-down considerations in general-purpose video encoding applications due to the lack of a particular or unique objective for the viewers. On the other hand, the main drawback of most bottom-up solutions lies in the huge amount of computational resources required to determine the Region Of Interest (ROI) and, consequently, the perceptually optimal resource allocation. Concerning this second issue, many simplified solutions make use of the connection between gaze and tracking of moving objects, identifying high-motion areas as the ROI and leaving aside other considerations. Priority measurement is consequently obtained by means of a certain motion activity characterization.

Although several works have made use of both spatial and temporal models for perceptual coding, resource allocation is typically performed by variating the Quantization Parameter (QP) on a macroblock basis (for example in [1] or [2]), which usually entails visible blocking artifacts for low bit rate applications. To solve this problem, some works have been proposed that filter the input sequence in order to remove high-frequency non-relevant information for bit rate reduction. For example, in [4], a bilateral pre-filtering is proposed that improves visual quality showing that a proper selection of filters allows for simplifying the sequence and reducing the average QP for encoding each frame, thus reducing undesirable visual artifacts.

Bilateral filtering has also been employed together with ROI estimation for the selection of the filter strength. In [5] an

activity map is obtained by measuring differences between current and previous frames and each pixel is filtered using one of two possible filter parameter sets, depending on the amount of activity measured. A more sophisticated approach is employed in [6] for the ROI characterization where the estimated saliency map controls the bilateral filter parameters. Both approaches aim to combine texture and ROI features with a filtering simplification stage prior to the encoding process, but the first one employs a severely restricted activity measure and exhibits poor filter adaptation, while the second one entails the use of too complex algorithms for determining the ROI. Additionally, neither of them seem to manage the problem of sequences with camera motion, which makes it more difficult to properly estimate the ROI.

The objective of this paper is to obtain a simple and effective method for perceptual resource allocation by filtering the sequence prior to the encoding process. This simplification is performed by a structure-preserving and priority-aware bilateral filtering with a moderately complex temporal activity analysis, which allows us to estimate the ROI. Additionally, a Camera Motion Compensation (CMC) algorithm is employed to correct eventual errors in this activity estimation due to camera motion.

The paper is organized as follows. Our proposal is described in Section II, which is organized in three sub-sections: bilateral filtering as the key technique for structural information preservation, the use of non-local approaches for temporal activity estimation and the addition of the CMC algorithm. Experimental results are shown and discussed in Section III. Finally, some conclusions are drawn in Section IV.

II. Perceptually-aware preprocessing for improved bit allocation

A. Texture-adaptive filtering

As mentioned in Section I, bilateral filters have been proved to be a good choice to pre-process video sequences due to their edge-preserving properties. These filters entail the use of two different operations, namely domain and range filters, which are combined as follows:

$$I_f(p) = \frac{1}{Z}\sum_{q\in\Omega} w_d(p,q,\sigma_d)w_r(p,q,\sigma_r)I(q), \quad (1)$$

where p is the current pixel position, with intensity value $I(p)$, and q is a neighboring pixel position belonging to a square region Ω around p. In our proposal, Ω is fixed to ± 3 pixels in both horizontal and vertical directions in order to keep the complexity low. Z is a normalization parameter and w_d and w_r are pixel-wise weighting functions for the domain and range filters, with parameters σ_d and σ_r, respectively. The first one is a Gaussian spatial filtering, while the second one operates on the intensity level domain. The corresponding weighting functions are defined as follows:

$$w_d(p,q,\sigma_s) = \exp\Big(-\frac{(p-q)^2}{2\sigma_d^2}\Big), \quad (2)$$

$$w_r(p,q,\sigma_r) = \exp\Big(-\frac{(I(p)-I(q))^2}{2\sigma_r^2}\Big), \quad (3)$$

As can be seen in Eq. 3, the range filter weight depends on the similarity between pixel values. Then, pixels belonging to object edges (i.e., with values different from their neighborhood) are not severely averaged and edge sharpness is preserved.

B. Motion-adaptive filtering

Bilateral filtering preserves edges automatically, enabling its use as an adaptive perceptually-aware filtering, but a second step incorporating ROI-related information is needed for a complete solution, obtaining the so-called Motion-Adaptive Bilateral Filter (MABF). In order to keep the algorithm as simple as possible, we employ an approach similar to that in [5], in which the frame difference is used as an estimate of motion activity. Our proposal makes use of a temporal analysis stage inspired by the neighbor-based distance suggested in [7] for non-local filtering, but employing it not for actually filtering but for modulating bilateral filter weights. This stage compares each pixel neighborhood with the co-situated region in previous frame, in order to accurately determine whether the pixel motion is high or low by means of a stationarity measure:

$$w_s(p) = \exp\Big(-\frac{\sum_{p\in\Omega}(I_n(p)-I_{n-1}(p))^2}{h^2}\Big), \quad (4)$$

where I_n and I_{n-1} are, respectively, the current and previous frames and h is an experimental parameter. The region Ω is the same as in Eq. (1).

The pixel stationarity index is used to modulate the filter parameters in order to increase the filter strength in those areas belonging to static regions, i.e. non-priority regions ($w_s(p)$ near 1), and to decrease the filtering strength in those pixels belonging to dynamic regions ($w_s(p)$ down to 0), those hopefully related to the ROI. For our proposal, the domain filter parameter σ_d is fixed to a low value in order to obtain a certain degree of simplification without the concurrence of undesirable blurring effects, while the range filter parameter is modulated by the stationarity measure as follows:

$$\sigma_r(p) = \sigma_{r0}\exp\Big(-\frac{(w_s(p)-1)^2}{\alpha}\Big), \quad (5)$$

where α is an experimental parameter and σ_{r0} represents the maximum amount of filtering.

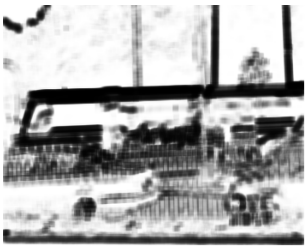
C. Camera Motion Compensation

In order to obtain a general purpose solution, our initial proposal has been adapted to compensate camera motion. The so-called Camera Motion Compensated (CMC) MABF makes use of a camera motion model for modifying the stationarity measure by spatially shifting the comparison described in Eq. (4) to the corresponding region according to the estimated camera motion:

$$w_s^{'}(p) = \exp\Big(-\frac{\sum_{p\in\Omega}(I_n(p)-I_{n-1}(p-mv_c(p)))^2}{h^2}\Big), \quad (6)$$

(a)

(b)



(c)

Fig. 1. (a) Frame #80 of *Bus* sequence; (b) stationarity map without CMC; (c) stationarity map with CMC. Brighter points mean higher stationarity values.

with $mv_c(p)$ being the camera motion vector for the current position p. When used to modulate the filter strength, the modified pixel stationarity measure $w_s^{'}(p)$, obtains the same results as Eq. (4) for sequences with no camera motion, but when camera motion is present, the behavior is quite different and those regions with no motion with respect to the camera are classified as ROI and, therefore, preserved.

The same technique described in [8] for camera motion estimation has been implemented, which entails the use of a hierarchical motion estimation and the application of well-known camera motion models. Nevertheless, the method to obtain the camera motion map is out of the scope of this paper, so any other approach could be employed.

III. Experiments and Results

The proposed algorithm has been tested on several CIF sequences. An stationary map has been obtained by means of Eq. (6) for each sequence, which has been then pre-filtered and encoded using the Joint Model (JM) H.264 reference software version JM12.2, available in [9], at a low bit rate of 128 kbps. The first 100 frames of each sequence were encoded using an IP..P pattern and one Intra picture every 26 frames. The algorithm parameters for filter adaptation have been chosen experimentally to avoid undesirable distortion according to subjective criteria. In particular, we have used $\sigma_d = 3$, $\alpha = 0.6$, $h = 10$ and $\sigma_{r0} = 10$.

In order to demonstrate the need of a CMC algorithm, some experiments have been performed for assessing the proposed stationarity measure. As an illustrative example, Fig. 1 shows that in sequences exhibiting camera motion, the motion activity in the background is higher and, as shown in Fig. 1(b), low stationarity values are obtained by the non-compensated camera motion algorithm, wrongly considering these regions as the ROI, while foreground objects (like the bus), which have little or no motion at all, appear in this case as static and non-relevant regions. On the other hand, the use of CMC techniques solves this problem with a proper stationarity map (see Fig. 1(c)). Some examples of frame captures are shown in Figures 2, 3 and 4, where some blurring is introduced as a tradeoff with blocking artifacts, reducing the average QP in about 0.5 units and notably improving quality in the ROI, especially in the head of the tennis player in Fig. 3 and in the player numbers in Fig. 4. Additionally, some examples of reconstructed sequences are available in [10].

IV. Conclusion

Our proposal of a motion-adaptive bilateral pre-filtering achieves a good performance in terms of perceptual quality improvement, while remaining simple enough to be used in almost every video compression scenario, by means of enhancing the properties of structural information preservation owing to bilateral filtering with the addition of a non-local temporal analysis for ROI estimation. CMC-MABF also benefits from the fact that pre-filtering approaches are not dependent on the actual video encoder or the coding standard employed for video compression, so the bit allocation is carried out without QP variations, which would lead to undesirable artifacts.

References

[1] K. Minoo and T. Nguyen, "Perceptual video coding with h.264," in *Signals, Systems and Computers, Conference Record of the Thirty-Ninth Asilomar Conference on*, 2005, pp. 741–745.

[2] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *Multimedia, IEEE Transactions on*, vol. 8, no. 1, pp. 11–18, 2006.

[3] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on*, vol. 13, no. 10, pp. 1304–1318, 2004.

[4] M.-Q. Li and Z.-Q. Xu, "An adaptive preprocessing algorithm for low bitrate video coding," *Journal of Zhejiang University - Science A*, vol. 7, no. 12, pp. 2057–2062, 2006.

[5] J.-H. Kim, J. W. Lee, R.-H. Park, and M.-H. Park, "Adaptive edge-preserving smoothing and detail enhancement for video preprocessing of h.263," in *Consumer Electronics (ICCE), International Conference on*, 2010, pp. 337–338.

[6] S.-P. Lu and S.-H. Zhang, "Saliency-based fidelity adaptation preprocessing for video coding," *Journal of Computer Science and Technology*, vol. 26, no. 1, pp. 195–202, 2011.

[7] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, Jun. 2005, pp. 60–65.

[8] A. Mejia-Ocana, M. de Frutos-Lopez, S. Sanz-Rodriguez, O. del Ama-Esteban, C. Pelaez-Moreno, and F. Diaz-de Maria, "Low-complexity motion-based saliency map estimation for perceptual video coding," in *Telecommunications (CONATEL), 2nd National Conference on*, 2011.

[9] "JM 10.2 [Online], http://iphome.hhi.de/suehring/tml/download/old_jm/."

[10] "Encoding test results." [Online]. Available: http://www.tsc.uc3m.es/~sescalona/PCS2012/

(a) (b)

Fig. 2. Frame #22 of *Bus* encoded at 128 kbps without filtering (a) and with CMC-MABF (b).

(a) (b)

Fig. 3. Frame #11 of *Stefan* sequence encoded at 128 kbps without filtering (a) and with CMC-MABF (b).

(a) (b)

Fig. 4. Frame #18 of *Football* sequence encoded at 128 kbps without filtering (a) and with CMC-MABF (b).